Guide - Econometrics II

Guilhermo Pastore and Lucas Oliveira

December 2025

Contents

1 Maximum Likelihood

1.1 What is likelihood?

Etymologically, likelihood refers to similarity to reality. Statistically, finding a maximum likelihood (ML) estimator means choosing the parameter values under which the observed data are most likely to have been generated. To understand this idea, let us begin with a simple example.

Suppose we observe the following random sample of five individuals' heights, measured in centimeters:

$$(170, 171, 172, 173, 174)$$

We would like to infer the true mean height in the population based on this sample. If you had to guess the population mean, what would be a reasonable guess?

At this point, it is important to emphasize that the observed sample is treated as fixed. What varies is the value of the parameter we are trying to infer, in this case, the population mean. Different guesses for the mean imply different explanations for how likely it is that this particular sample was observed.

For instance, if the true mean were 150, observing heights as large as 170 or 174 would seem quite unusual. Likewise, if the true mean were 190, it would also be strange to observe such relatively low values. In both cases, the probability of observing this specific sample would be small.

Implicitly, this reasoning assumes that heights are generated by a probability distribution indexed by a mean parameter, such as a normal distribution. Different values of the mean assign different probabilities to the same observed data.

Intuitively, we therefore want to choose the mean that makes the observed sample look least "weird," or, equivalently, most plausible under the assumed distribution. In this example, that value turns out to be 172, the sample mean.

Maximum likelihood estimation formalizes exactly this idea: it selects the parameter value that maximizes the likelihood of observing the given sample.

1.2 ML formalization

Let $(y_1, \ldots, y_n)$ be a random sample of independent and identically distributed random variables. Assume that each observation has probability density function $f(y_i|\theta)$, where $\theta \in \Theta \subset \mathbb{R}^k$ is a vector of unknown parameters.

The joint density of the sample, conditional on θ, is given by

$$f(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

Once a realization $y = (y_1, \ldots, y_n)$ is observed, the same expression can be viewed as a function of the parameter vector θ. This function is called the likelihood function and is denoted by

$$L(\theta|y) = \prod_{i=1}^{n} f(y_i|\theta)$$

The maximum likelihood estimator $\hat{\theta}_{ML}$ is defined as any value of θ that maximizes the likelihood function:

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta|y)$$

1.3 Log-likelihood

As we saw in 1.2, the likelihood function is defined as a product of individual densities. While this representation is conceptually useful, working directly with products can be inconvenient, especially for optimization and analytical derivations.

For this reason, it is standard practice to consider the logarithm of the likelihood function, known as the log-likelihood:

$$\ell(\theta|y) = \log\,[L(\theta|y)] = \sum_{i=1}^{n} \log\,[f(y_i|\theta)]$$

Taking logarithms does not change the location of the maximizer, since the logarithm is a strictly increasing transformation. Therefore,

$$\arg\max_{\theta \in \Theta} L(\theta|y) = \arg\max_{\theta \in \Theta} \ell(\theta|y)$$

1.4 Example: Normal distribution with known variance

To illustrate maximum likelihood estimation in a concrete setting, consider again the example of individual heights. Assume that the observations $y_1, \dots, y_n$ re independently drawn from a normal distribution with unknown mean μ and known variance σ^2:

$$y_i \sim \mathcal{N}(\mu, \sigma^2)$$

The probability density function of a single observation is

$$f(y_i|\mu) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}$$

The log-likelihood function for the sample is therefore given by

$$\begin{aligned}\ell(\mu|y) &= \sum_{i=1}^{n} \log\,[f(y_i|\mu)] = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\sigma^2\pi}}\right) - \frac{(y_i-\mu)^2}{2\sigma^2} \\ &= \sum_{i=1}^{n} -\frac{1}{2}\log(2\sigma^2\pi) - \frac{(y_i-\mu)^2}{2\sigma^2} \\ &= -\frac{n}{2}\log(2\sigma^2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\end{aligned}$$

To find the estimator we calculate

$$\begin{aligned}\frac{\partial\ell(\mu|\theta)}{\partial\mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) = 0 \\ &\Rightarrow \sum_{i=1}^{n}(y_i-\mu) = 0 \\ &\Rightarrow \sum_{i=1}^{n} y_i - n\mu = 0 \\ &\therefore \hat{\mu}_{ML} = \frac{\sum_{i=1}^{n} y_i}{n}\end{aligned}$$

which is simply the sample mean.

This result shows that, under normality, the maximum likelihood estimator of the mean coincides with the intuitive choice discussed earlier: the value of μ that makes the observed data most plausible is the sample average.

1.5 Score Function and FOC's

Let $\ell(\theta|y)$ denote the log-likelihood function associated with a random sample $y = (y_1, \ldots, y_n)$ The score function is defined as the gradient of the log-likelihood with respect to the parameter vector θ:

$$g(\theta|y) = \nabla_\theta \ell(\theta|y) = \begin{pmatrix} \frac{\partial\ell(\theta|y)}{\partial\theta_1} \\ \vdots \\ \frac{\partial\ell(\theta|y)}{\partial\theta_k} \end{pmatrix}$$

Knowing that $\ell(\theta|y) = \sum_{i=1}^{n} \log\,[f(y_i|\theta)]$, for each $j = 1, \ldots, k$ we can see that

$$\frac{\ell(\theta|y)}{\partial\theta_j} = \frac{\partial}{\partial\theta_j}\left(\sum_{i=1}^{n} \log\,[f(y_i|\theta)]\right) = \sum_{i=1}^{n} \frac{\partial \log\,[f(y_i|\theta)]}{\partial\theta_j}$$

Stacking the derivatives:

$$g(\theta|y) = \sum_{i=1}^{n} \begin{pmatrix} \frac{\partial \log\ [f(y_i|\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log\ [f(y_i|\theta)}{\partial \theta_k} \end{pmatrix} = \sum_{i=1}^{n} g_i(\theta|y)$$

1.6 Expectation of the score

Supposing that the true value of the parameter is θ_0, we now show that, under standard regularity conditions

$$\mathbb{E}[g(\theta_0|y)] = \mathbf{0}$$

Proof. By linearity of expectation and the independence of the observations,

$$\mathbb{E}[g(\theta_0|y)] = \sum_{i=1}^{n} \mathbb{E}[\nabla_\theta \log\ f(y_i|\theta_0)]$$

Consider a generic term in the sum. Under regularity conditions that allow differentiation under the integral sign,

$$\mathbb{E}[\nabla_\theta \log\ f(y_i|\theta_0)] = \int \nabla_\theta \log\ f(y|\theta_0) f(y|\theta_0)\, dy$$

Using the identity

$$\nabla_\theta \log\ f(y|\theta_0) = \frac{\nabla_\theta f(y|\theta_0)}{f(y|\theta_0)}$$

we obtain

$$\int \nabla_\theta \log\ f(y|\theta_0) f(y|\theta_0)\, dy = \int \nabla_\theta f(y|\theta_0)\, dy$$

Interchanging differentiation and integration,

$$\int \nabla_\theta f(y|\theta_0)\, dy = \nabla_{\theta_0} \int\ f(y|\theta_0)\, dy$$

Since $f(y|\theta_0)$ is a pdf,

$$\int\ f(y|\theta_0)\, dy = 1 \Rightarrow \nabla_{\theta_0} \int\ f(y|\theta_0)\, dy = \nabla_{\theta_0} 1 = \mathbf{0}$$

Therefore,

$$\mathbb{E}[\nabla_\theta \log f(y_i|\theta)] = \mathbf{0}$$

summing over i, we conclude that

$$\mathbb{E}[g(\theta_0|y)] = \mathbf{0}$$

□

1.7 Hessian matrix and second-order conditions

The Hessian matrix of the log-likelihood is defined as the matrix of second-order partial derivatives with respect to the parameter vector θ:

$$H(\theta|y) = \nabla^2_\theta \ell(\theta|y) = \frac{\partial^2 \ell(\theta|y)}{\partial\theta\partial\theta'}$$

The Hessian summarizes the local curvature of the log-likelihood function. While the score indicates the direction of steepest ascent, the Hessian determines whether a stationary point corresponds to a maximum, minimum, or saddle point.

Under standard regularity conditions, if $\hat{\theta}_{ML}$ is an interior solution to the maximum likelihood problem, it must satisfy the first-order condition

$$g(\hat{\theta}_{ML}|y) = \mathbf{0},$$

together with the second-order condition

$$H(\hat{\theta}_{ML}|y) \prec 0,$$

that is, the Hessian evaluated at $\hat{\theta}_{ML}$ must be negative definite.

Negative definiteness of the Hessian ensures that the log-likelihood is locally concave around $\hat{\theta}_{ML}$ implying that the stationary point is a local maximum. In many likelihood-based models of interest, the log-likelihood is globally concave in θ, in which case the second-order condition guarantees that the maximum likelihood estimator is unique.

1.8 Fisher information matrix

The Hessian matrix introduced in the previous section characterizes the curvature of the log-likelihood function for a given sample. To study the statistical properties of the maximum likelihood estimator, it is useful to consider the expected curvature of the log-likelihood under the true data-generating process.

The Fisher information matrix is defined as

$$\mathcal{I}(\theta) = -\mathbb{E}[H(\theta|y)] = -\mathbb{E}[\nabla^2_\theta \ell(\theta|y)]$$

where the expectation is taken with respect to the distribution $f(y|\theta)$.

Now, we show that under standard regularity conditions, the Fisher information matrix admits an equivalent representation in terms of the score function:

$$\mathcal{I}(\theta) = \mathbb{E}[g(y|\theta)g(y|\theta)']$$

Proof. We have already shown that $\mathbb{E}[g_i(\theta_0)] = 0$. Now, let's look at the score variance:

$$\text{Var}[g_i(\theta_0)] = \mathbb{E}[g_i(\theta_0)g_i(\theta_0)'] - \underbrace{\mathbb{E}[g_i(\theta_0)]\mathbb{E}[g_i(\theta_0)]'}_{=0}$$

Therefore,

$$\text{Var}[g_i(\theta_0)] = \mathbb{E}[g_i(\theta_0)g_i(\theta_0)']$$

From earlier we know that:

$$g_i(\theta) = \nabla_\theta \text{log}[f_i(y_i|\theta)] = \frac{\nabla_\theta f_i(y_i|\theta)}{f_i(y_i|\theta)}$$
$$\Rightarrow g_i g_i' = \frac{(\nabla_\theta f)(\nabla_\theta f)'}{f^2}$$

Calculating the score derivative:

$$H_i(\theta) = \nabla_\theta g_i = \nabla_\theta \left(\frac{\nabla_\theta f}{f} \right)$$

By the quotient rule:

$$H_i(\theta) = \frac{\nabla_\theta^2 f}{f} - \frac{(\nabla_\theta f)(\nabla_\theta f)'}{f^2}$$
$$\Rightarrow -H_i(\theta) = \frac{(\nabla_\theta f)(\nabla_\theta f)'}{f^2} - \frac{\nabla_\theta^2 f}{f}$$

Taking the expectation, we obtain two terms. Looking at the first one, we acknowledge that:

$$\mathbb{E}\left[\frac{(\nabla_\theta f)(\nabla_\theta f)'}{f^2}\right] = \mathbb{E}[g_i g_i']$$

The second term is:

$$\mathbb{E}\left[\frac{\nabla_\theta^2 f}{f}\right] = \int \nabla_\theta^2 f(y|\theta)\, dy = \nabla_\theta^2 \int f(y|\theta)\, dy = \nabla_\theta^2 1 = 0$$

Therefore,

$$-\mathbb{E}[H_i(\theta)] = \mathbb{E}[g_i g_i']$$

For the sample

$$g(\theta) = \sum_{i=1}^n g_i(\theta), \quad H(\theta) = \sum_{i=1}^n H_i(\theta)$$

Knowing that g_i are independent terms and have mean equal to zero:

$$\mathbb{E}[gg'] = \sum_{i=1}^n \mathbb{E}[g_i g_i'] = -\sum_{i=1}^n \mathbb{E}[H_i] = -\mathbb{E}[H]$$

Therefore

$$\mathcal{I}(\theta) = \mathbb{E}[g(y|\theta)g(y|\theta)'] = -\mathbb{E}[H(\theta)]$$

□

This equivalence highlights an important interpretation of the Fisher information. While the Hessian measures the local curvature of the log-likelihood for a specific sample, the Fisher information measures the average curvature, or equivalently, the average squared magnitude of the score, under the true model.

Intuitively, the Fisher information captures how sensitive the likelihood function is to changes in the parameter vector θ. A large Fisher information matrix indicates that small changes in the parameter values lead to large changes in the likelihood, implying that the parameter can be estimated with greater precision.

1.9 Asymptotic normality

We now derive the asymptotic distribution of the maximum likelihood estimator using the properties of the score function and the Hessian matrix established in the previous sections.

Proof. Consider a first-order Taylor expansion of the score function around the true parameter value θ_0:

$$g(\hat{\theta}_{ML}|y) = g(\theta_0|y) + H(\bar{H}|y)(\hat{\theta}_{ML} - \theta_0)$$

where $H(\theta|y) = \nabla^2_\theta \ell(\theta|y)$ is the Hessian matrix, and $\bar{H}$ lies between $\hat{\theta}_{ML}$ and θ_0.

Since $g(\hat{\theta}_{ML}|y) = \mathbf{0}$, we can rearrange the expression to obtain

$$(\hat{\theta}_{ML} - \theta_0) = -H(\bar{\theta}|y)^{-1} g(\theta_0|y)$$

Multiplying both sides by $\sqrt{n}$

$$\begin{aligned}\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = -\sqrt{n}H(\bar{\theta}|y)^{-1}g(\theta_0|y) &= -\sqrt{n}\frac{n}{n}H(\bar{\theta}|y)^{-1}g(\theta_0|y)\\ &= -\sqrt{n}\left(\frac{1}{n}H(\bar{\theta}|y)\right)^{-1}\frac{1}{n}g(\theta_0|y)\\ &= -\left(\frac{1}{n}H(\bar{\theta}|y)\right)^{-1}\frac{1}{\sqrt{n}}g(\theta_0|y)\end{aligned}$$

The score evaluated at the true parameter value can be written as

$$g(\theta_0|y) = \sum_{i=1}^{n} g_i(\theta_0)$$

where $g_i(\theta_0) = \nabla_\theta \log[f(y_i|\theta)]$. Under standard regularity conditions,

$$\mathbb{E}[g_i(\theta_0)] = \mathbf{0}, \quad \mathrm{Var}[g_i(\theta_0)] = \mathcal{I}(\theta_0)$$

By the Central Limit Theorem,

$$\frac{1}{\sqrt{n}} g(\theta_0|y) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}(\theta_0))$$

The normalized Hessian can be expressed as

$$\frac{1}{n} H(\bar{\theta}|y) = \frac{1}{n} \sum_{i=1}^{n} H_i(\bar{\theta}|y)$$

Under consistency of $\hat{\theta}_{ML}$ and regularity conditions ensuring uniform convergence,

$$\frac{1}{n} H(\bar{\theta}|y) \xrightarrow{p} \mathbb{E}[H_i(\theta_0)] = -\mathcal{I}(\theta_0)$$

Consequently by the Continuous Map Theorem,

$$\left(\frac{1}{n} H(\bar{\theta}|y)\right)^{-1} \xrightarrow{p} -\mathcal{I}(\theta_0)^{-1}$$

Combining the convergence results above and applying Slutsky's theorem, we obtain

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}(\theta_0)^{-1})$$

□

2 Discrete Choice

Discrete choice models often arise when the dependent variable takes only two possible values, typically coded as 0 or 1. These models are widely used in applied econometrics and appear in two closely related contexts.

First, the analyst may observe a binary outcome and a set of covariates, and wish to model the relationship between them. Second, the binary outcome may arise as a coarsened observation of an underlying economic decision.

Despite these different interpretations, the econometric tools used in both cases are largely the same.

2.1 The Binary Choice Problem

Let $y_i \in \{0, 1\}$ denote a binary outcome for individual i and let x_i be a vector of observed covariates.

The object of interest is the conditional probability

$$\mathbb{P}(y_i = 1|x_i) = F(x_i'\beta)$$
$$\mathbb{P}(y_i = 0|x_i) = 1 - F(x_i'\beta)$$

where $F(\cdot)$ is a function mapping the real line into the unit interval.

2.2 The Linear Probability Model

A natural starting point is the linear probability model (LPM),

$$F(x_i'\beta) = x_i'\beta.$$

Because

$$\mathbb{E}[y_i|x_i] = \mathbb{P}(y_i = 1|x_i),$$

this leads to the regression equation

$$y_i = x_i'\beta + \varepsilon_i$$

Although simple, the LPM suffers from important drawbacks:

- predicted probabilities may lie outside $[0, 1]$
- the error term is heteroskedastic
- the model imposes constant marginal effects

2.3 Link Functions and Nonlinear Probability Models

A desirable probability model should satisfy:

$$\lim_{x_i'\beta \to +\infty} \mathbb{P}(y_i = 1|x_i) = 1, \quad \lim_{x_i'\beta \to -\infty} \mathbb{P}(y_i = 0|x_i) = 0$$

Any continuous cumulative distribution function defined on the real line satisfies these conditions. Two choices are particularly common.

2.4 Probit

In the probit model, the probability of success is given by

$$\mathbb{P}(y_i = 1|x_i) = \Phi(x_i'\beta)$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

The probit model assumes a symmetric response of probabilities to changes in the index $x_i'\beta$, with thinner tails than the logistic distribution.

2.5 Logit

In the logit model, the probability of success is

$$\mathbb{P}(y_i = 1|x_i) = \Lambda(x_i'\beta) = \frac{e^{x'_i\beta}}{1 + e^{x'_i\beta}}$$

where $\Lambda(\cdot)$ is the logistic cumulative distribution function.

The logit model is often preferred for its analytical convenience and closed-form expressions for derivatives.

2.6 Marginal Effects

In binary choice models, the conditional expectation of the dependent variable is equal to the probability of success:

$$\mathbb{E}[y_i|x_i] = \mathbb{P}(y_i = 1|x_i) = F(x_i'\beta)$$

Because $F(\cdot)$ is a nonlinear function, the coefficients β do not directly represent marginal effects, unlike in linear regression models. As a result, interpretation requires special care.

2.6.1 General Form

For a continuous covariate x_{ik}, the marginal effect on the probability of success is defined as

$$\frac{\partial \mathbb{E}[y_i|x_i]}{\partial x_{ik}} = \frac{dF(x_i'\beta)}{d(x_i'\beta)} \cdot \beta_k = f(x_i'\beta) \cdot \beta_k$$

where $f(\cdot) = F'(\cdot)$ is the corresponding probability density function, and β_k is the coefficient associated with x_{ik}

Thus, marginal effects depend on the value of the covariates through $x_i'\beta$ and the functional form of the link function.

This feature distinguishes nonlinear probability models from the linear probability model, in which marginal effects are constant.

2.6.2 ME in Probit

In the probit model,

$$F(x_i'\beta) = \Phi(x_i'\beta)$$

and the derivative of the CDF is the standard normal density:

$$f(x_i'\beta) = \phi(x_i'\beta)$$

The marginal effect is therefore

$$\frac{\partial \mathbb{E}[y_i|x_i]}{\partial x_{ik}} = \phi(x_i'\beta)\beta_k$$

2.6.3 ME in Logit

In the logit model,

$$F(x_i'\beta) = \Lambda(x_i'\beta)$$

where

$$\Lambda(z) = \frac{e^z}{1+e^z}$$

the derivative of the logistic CDF is:

$$\frac{d\Lambda(z)}{dz} = \Lambda(z)[1-\Lambda(z)]$$

Hence, the marginal effect in the logit model is

$$\frac{\partial \mathbb{E}[y_i|x_i]}{\partial x_{ik}} = \Lambda(x_i'\beta)[1-\Lambda(x_i'\beta)]\beta_k$$

2.7 Evaluating Marginal Effects

Because marginal effects depend on x_i, they must be evaluated at specific values of the covariates. Two common approaches are used in practice.

2.7.1 Marginal Effects at the Mean (MEM)

One approach is to evaluate marginal effects at the sample means:

$$\left.\frac{\partial \mathbb{E}[y|x]}{\partial x}\right|_{x=\bar{x}}.$$

This method is simple but may be misleading when the mean covariate vector does not correspond to a representative individual.

2.7.2 Average Marginal Effects (AME)

An alternative is to compute marginal effects for each observation and then average:

$$AME_k = \frac{1}{N}\sum_{i=1}^{N} f(x_i'\hat{\beta})\hat{\beta}_k$$

Under regularity conditions, both approaches are asymptotically equivalent. In finite samples, however, average marginal effects are generally preferred.

2.7.3 Marginal Effects for Binary Regressors

When a covariate x_{ik} is binary, derivatives are no longer meaningful. Instead, marginal effects are defined as discrete changes:

$$\Delta_k = \mathbb{P}(y_i = 1|x_{ik} = 1, x_{i,-k}) - \mathbb{P}(y_i = 1|x_{ik} = 0, x_{i,-k})$$

This definition applies to both probit and logit models and is evaluated either at the mean of the remaining covariates or averaged across observations.

2.8 Latent Index Models and Threshold Crossing

Firstly, let's consider an example in which an individual chooses whether or not to participate in the labor force. Let

$$y_i = \begin{cases} 1, & \text{if individual } i \text{ works or actively searches for work} \\ 0, & \text{otherwise} \end{cases}$$

This observed outcome is binary. However, economic theory suggests that the underlying decision is not binary.

Suppose individual i compares the benefits and costs of participating in the labor force. Define a latent variable y_i^*, representing the net benefit of participation, which may depend on observable characteristics such as age, education, marital status, number of children, among others.

We can model this benefit as

$$y_i^* = x_i'\beta + \varepsilon_i,$$

where $x_i'\beta$ captures observed determinants of labor supply and ε_i captures unobserved factors, such as preferences for leisure or family responsibilities.

The econometrician does not observe y_i^*. Instead, the observed binary outcome is generated by a threshold-crossing rule:

$$y_i = \begin{cases} 1, & \text{if } \; y_i^* > 0 \\ 0, & \text{if } \; y_i^* \leq 0 \end{cases}$$

Thus, participation occurs whenever the net benefit of working is positive. This threshold-crossing rule converts a continuous latent variable into a discrete observed outcome.

2.8.1 From the Latent Variable to Choice Probabilities

Using the threshold-crossing rule, the probability of observing $y_i = 1$ is

$$\mathbb{P}(y_i = 1|x_i) = \mathbb{P}(\varepsilon_i > -x_i'\beta|x_i)$$

Let $F(\cdot)$ denote the cumulative distribution function of ε_i. Then,

$$\mathbb{P}(y_i = 1|x_i) = F(x_i'\beta).$$

This expression coincides exactly with the probability specification introduced in Subsection 2.1. Thus, probit and logit models can be interpreted as latent index models with different assumptions on the distribution of the unobserved component.

2.8.2 Probit and Logit as Latent Variable Models

Probit

$$\varepsilon_i \sim \mathcal{N}(0,1) \Rightarrow \mathbb{P}(y_i = 1|x_i) = \Phi(x_i'\beta).$$

Logit

$$\varepsilon_i \sim \text{Logistic}(0, \pi^2/3) \Rightarrow \mathbb{P}(y_i = 1|x_i) = \Lambda(x_i'\beta)$$

Both models differ only in the assumed distribution of the latent error term.

2.8.3 Identification and Scale Normalization

The latent variable y_i^* is not observed, and only its sign matters for the observed outcome. As a result, the scale of the model is not identified.

Suppose instead that

$$y_i^* = x_i'\beta + \sigma\varepsilon_i$$

where $\sigma > 0$ is an unknown scale parameter. Dividing both sides by σ,

$$\frac{y_i^*}{\sigma} = x_i'\left(\frac{\beta}{\sigma}\right) + \varepsilon_i$$

Because the sign of y_i^* is unchanged by this transformation, the parameters β and σ cannot be separately identified. Consequently, the variance of ε_i must be normalized.

2.8.4 Threshold Normalization

The threshold at which the latent variable is converted into the observed outcome is also subject to normalization.

More generally, suppose the observed rule is

$$y_i^* = 1 \;\; \text{if} \;\; y_i^* > \alpha,$$

where α is an unknown threshold. If the model includes an intercept term, this threshold is not separately identified from the constant.

Therefore, it is without loss of generality to normalize the threshold to zero and include a constant in x_i.

2.9 MLE Estimations for Binary Choice Models